Unsupervised Semantic Segmentation with Pose Prior

Max Midwinter, Zaid Abbas Al-Sabbag, Chul Min Yeum*

[1]Department of Civil and Environmental Engineering, Faculty of Engineering, University of Waterloo, Waterloo, Ontario, Canada, N2L 3G1
Email : mxxmidwi@uwaterloo.ca

ABSTRACT: The most popular computer vision-aided visual inspection techniques in civil engineering are single image, deep neural networks, trained in a supervised learning scheme. With deeper convolutional networks and vision transformers, the need for training data will exceed the supply of and researchers' ability to label training images. In this work, the authors will explore whether poses can be a viable inductive bias to produce a semantic segmentation with a shallow unsupervised convolutional neural network. In this work the authors utilize commercial augmented reality devices, Microsoft HoloLens 2, running simultaneous localization and mapping, to collect images and global poses of spalling defects. An unsupervised semantic segmentation, based on, and novel stochastic refinement and outlier rejection algorithm are used to generate segmentation of spalling defects. The results of experiments are positive yielding an average mIOU of 86.25%. The authors believe this result and the emergence of pose trackers (e.g., LiDARs, depth sensors), has great potential to yield a synergetic relationship between deep learning and robotics.

KEYWORDS: Visual Inspection, Unsupervised Learning, Semantic Segmentation, Augmented Reality/Mixed Reality

1 INTRODUCTION

Structural inspections are a critical, routine, and labor-intensive task; the common bridge soffit inspection can be extremely involved, requiring specialized equipment such as mechanical lifts and booms. By exploiting developments in computer vision and artificial intelligence, researchers have proposed to automate structural inspections, which would improve inspection outcomes with a reduction in costs. A typical visual inspection involves the inspector studying the visual appearance of the infrastructure. The inspectors make notes of any defects in the infrastructure and if applicable make a measurement or visual estimate of the defect for records and repair estimates. In other words, a structural inspection involves classification, localization, and quantification of the defect. To this end, researchers have adopted various convolutional neural network (CNN) architectures with the capability to classify, detect and segment defects in inspection images.

Defect segmentations are synergetic with augmented reality inspection schemes as this would allow the inspector to quantify the defect in a real-world scale [2]. However, collecting data and creating segmentation labels is an extremely labor-intensive task. Additionally, in augmented reality (AR) enabled inspection, a defect will likely appear in the frame multiple times. If a traditional single image supervised segmentation method is used, it is not intuitive to determine the best segmentations from the multiple segmented instances. This would require the further intervention of the inspecting engineer.

In this work, the authors utilize the Microsoft HoloLens 2 (HL2) MR device to collect defect images, Azure 3D spatial mesh and poses, which are utilized by the proposed unsupervised semantics segmentation with pose prior (USP) method.

2 METHODOLOGY

2.1 Pre-processing

The pre-processing starts with a sampled set (or sequence) of images, with known poses, containing a region of interest (ROI) selected or detected. To start, all images are transformed or rectified to the reference frame view. As mentioned above, this is accomplished by exploiting homography between images. This is usually done with feature point correspondences; however, this may be unreliable in civil environments due to lack of features, poor lighting, etc.

In this study, 4 points (corners of a bounding box) are chosen by the HL2 user, around the defect to segment. HL2 points are ray casted and anchored to HL2's 3D spatial mesh [2]. Homography then can be found in the following way [3, 4]:

$$H_{ab} = R_a R_b^T - \frac{(-R_a R_b^T t_b + t_a) n^T}{d} \qquad (1)$$

The sequence of images is transformed to image frame b, where the inspector selected the bounding box.

2.2 Unsupervised Segmentation and Outlier Rejection

The preprocessed images are segmented independently using unsupervised differentiable feature clustering developed by Kim et. al. [1]. This technique utilizes a lightweight CNN composed of only three convolution blocks (2D convolution, ReLU activation, batch-normalization); and attempts to iteratively meet (i.e., minimize the loss from) three incompatible requirements: pixels of similar features should be assigned the same label, spatially continuous pixels should be assigned the same label, and the number of unique labels should be large.

The quality of the semantic segmentation can vary as the unsupervised segmentation method is nondeterministic and can be degraded by environmental factors such as low lighting, motion blur etc. Civil environments are particularly challenging for computer vision due to the relative lack of features of large concrete surfaces. Thus, it is necessary to reject incorrect segmentations (outliers) and prevent them from influencing the final segmentation result. Since there is no ground truth, the authors utilize a relative similarity metric; based on the assumption that in a high degree of freedom space, outliers are more different than inliers. To identify outliers, each segmentation result is compared to each other, with the structural similarity metric (SSIM), which returns a value between 0 (no similarity) and 1 (same image) [5].

2.3 Stochastic Consensus

The purpose of the stochastic consensus process is to generate seed points, from inlier segmentations, which will be used in the next iteration of unsupervised semantic segmentation and outlier rejection. Seed points are a form of pseudo-labelling, where high confidence predictions are used as supervision for training.

Since the generated segmentations are pseudo-labels, there is a chance that labels are incorrect. To ensure there is sufficient supervision and the impact of incorrect labels is minimized, the authors propose to sample a small percentage of pixel locations of the inlier segmentation masks, this is referred to as seed point coverage. If all inlier masks, at a pixel location, predict the same class then that location will become a seed point for that class. Found seed points are collated into the seed point mask, which is used by the unsupervised semantic segmentation to produce better segmentations.

3 EXPERIMENTS

The experiment was conducted at the Gardiner underpass located on Wickman Rd., Toronto, Ontario, Canada, which is a pedestrian-accessible road tunnel running under the Gardiner Expressway. The underpass contains spalling damage on the sides of the tunnel which are accessible on the ground level, see Figure 1.

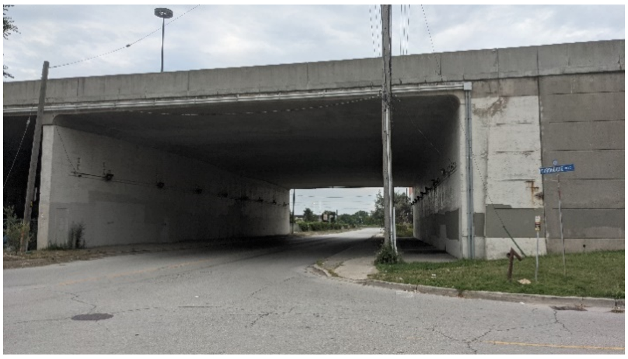

Figure 1. Gardiner Expressway Overpass at Wickman Rd.

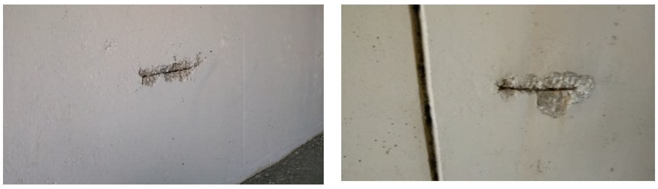

Figure 2. Defect 1 (left) and Defect 2 (right).

Defects present on the bridge pier are shown in Figure 2. And their ground truth segmentation and predicted segmentation are shown in Figure 3, below. The segmentation accuracy of the two defects is summarized in Table 1.

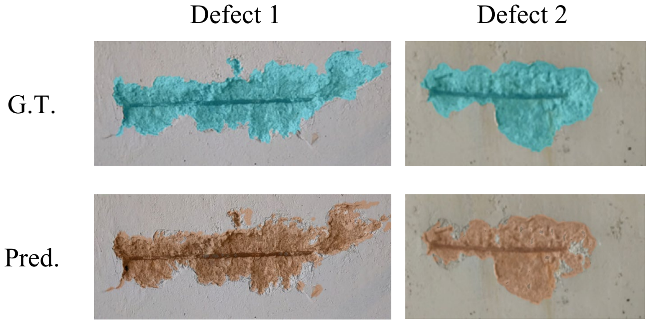

Figure 3. Segmentation Results Ground Truth (G.T.) verses Predicted.

Table 1. mIOU in percentage for Defect 1 and Defect 2.

Defect	mIOU (%)
Defect 1	84.8
Defect 2	87.7

4 CONCLUSION

The proposed methodology (USP) enables pixel-wise semantic segmentation, without training data, by leveraging multiple images with known poses. Pixel-wise segmentation allows the selection of the defect area, which is important for applications such as quantifying the change in defect size over multiple inspections.

While USP should be able to be extended to various inspection applications; the broader implication is that poses can be used to introduce inductive biases to a neural network, that can be exploited to create a new self-learning paradigm.

REFERENCES

[1] Kim, W., Kanezaki, A., & Tanaka, M. 2020. “Unsupervised learning of image segmentation based on differentiable feature clustering.” IEEE Transactions on Image Processing, 29, 8055-8068.

[2] Al-Sabbag, Z. A., Yeum, C. M., & Narasimhan, S. (2022). Interactive defect quantification through extended reality. Advanced Engineering Informatics, 51, 101473. https://doi.org/10.1016/j.aei.2021.101473

[3] Li, B., Heng, L., Lee, G. H., & Pollefeys, M. (2013, November). A 4-point algorithm for relative pose estimation of a calibrated camera with a known relative rotation angle. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1595-1601). IEEE.

[4] Malis, E., & Vargas, M. (2007). Deeper understanding of the homography decomposition for vision-based control.

[5] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.